UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Immune subtype in Colorectal Cancer: Molecular, functional characterization and clinical implications.

Máster Universitario en Bioinformática y Biología Computacional

Autor: CARRETERO PUCHE, CARLOS

Tutor: GOMEZ LOPEZ, GONZALO

Departamento de Bioquímica

FECHA: Septiembre 2019

ABSTRACT

Purpose: Cancer initiation and progression are the consequence of a complex interplay between cancer cells and the tumor microenvironment. Recently, a new global transcriptomic immune classification of solid tumors has identified six immune subtypes (ISs) (C1-C6) with distinct immunogenomic features significantly associated with clinical outcome. The aim of our study was to specifically characterize the ISs in colorectal cancer (CRC) patients, and to assess their interplay with the Consensus Molecular Subtypes (CMS).

Patients and Methods: CMS and ISs data, as well as clinical and molecular information of CRC patients were obtained from TCGA database27 (N=625). Immune cell populations (CIBERSORT, MCP-Counter methods), differential gene expression analyses and Gene Set Enrichment Analysis (GSEA) were also performed to characterize the immune subtypes in the global CRC population and by CMS subtype.

Results: Only 5 of the 6 ISs are present in CRC patients, with 2 predominant ISs: the C1 Wound Healing (77%) and the C2 IFN-γ Dominant (17%) subtypes. CMS1 showed the highest proportion of C2 (53%), while C1 was particularly dominant in CMS2 (91%). CMS3 had the highest representation of C3 Inflammatory (7%) and C4 Lymphocyte Depleted ISs (4%), while C6 TGF-β Dominant cases belonged exclusively to CMS4 (2.3%). The prognostic impact of ISs in CRC was substantially different from that reported for the global TCGA dataset, with best 5-year survival rates observed in C6 and C1 patients (100% and 65%, respectively), while C2 and C3 displayed the worst outcome (49% and 23%, respectively). C2 tumors had a high density of CD8, follicular helper T cells, regulatory T cells, dendritic cells and neutrophils, while C1 was enriched in plasma, CD4 T and activated mast cells. Accordingly, expression of several immunomodulatory genes, including immune-checkpoints (PDL-1, CTL4, LAG 3), was upregulated in C2 tumors. GSEA analysis revealed C2 was characterized by a high activation of the immune system, apoptosis and DNA repair, as well as mTOR signalling and oxidative phosphorylation, while C1 was more dependent of metabolic pathways such as glycolysis and pyruvate metabolism.

Conclusion: ISs identify distinct immune profiles within CMS subgroups with relevant clinical and biological implications and may therefore be a valuable tool to improve tailored therapeutic interventions in CRC patients.

ACKNOWLEDGMENTS

I would like to thank all member of my researcher group, the cancer digestive unit in Hospital 12 de Octubre, specially to Beatriz Soldevilla Navarro and Rocío García Carbonero who have leaded me the clinic part of the work. In addition, I would like to thank the all members of CNIO's Bioinformatics Unit, and specially to Fátima and my tutor Gonzalo Gomez Lopez who have always a good advice for you.

MOTIVATION

Colorectal cancer (CRC) is the second leading cause of cancer-related death worldwide and the first taken into account both genders, with 15,410 deaths in 2017 and an estimation of 44,937 new cases in Spain according to the Sociedad Española de Oncología Médica (https://seom.org). CRC clinical management is primarily based on tumor location and extent of the disease, with a very limited repertoire of molecular biomarkers to guide personalized patient care. CRC is, however, a heterogeneous disease with widely variable clinical outcomes, both in terms of prognosis and drug response. Mainly part of this clinical variability is due to tumor diversity that is also reflected at genomic level where two major groups can be identified with relevant predictive and prognostic implications: i) hypermutated tumors (mutation rates > 12xE06 bp), accounting for about 15% of colorectal carcinomas which are associated with a rich immune cell infiltration, favourable prognosis in early-stage disease, poor prognosis in the metastatic setting but a good response to immune checkpoint inhibition and ii) non-hypermutated tumors that are globally associated with a worse prognosis and a poor response to immunotherapy[1,2]. Despite the genomic heterogeneity in CRC, the only genomic biomarkers that are currently well established as standard guidelines are *RAS* mutation to avoid therapies with anti-*EGFR* drugs and MSI or deficient MMR to select patients for treatment with immune checkpoint inhibitors[1–5].

Molecular classification based on transcriptomics data is a tool that is taking importance at scientific and clinical level and provide new ways of clinical cancer research[6–8]. This technology combines information of several genes that are expressing, not only in tumor cells, but also in those that are part of microenvironment. It is well known that microenvironment plays an important role determining the patient prognosis and response to therapies. In this context, transcriptomics classification encompasses not only tumor

cell component, but also immune and stromal components of cancer. Thus, to make classifications based on intrinsic gene expression profiles more intimately linked to tumor phenotype and therefore more closely related to clinical behaviour[9–14]. In 2015, the international Colorectal Cancer Subtyping Consortium (CRCSC) proposed a unified transcriptomics classification that identified four biologically distinct Consensus Molecular Subtypes (CMS)[15]. Although CMS classification show substantial differences at abundances of immune populations and expression of immune related genes, this classification is based on a general transcriptomics data. In contrast with CMS classification, Thorsson and colleagues have developed a clustering system based on transcriptomics data but focussing only on genes related to immune system[16]. Nowadays, the study of tumors in an immune perspective have acquired high clinical relevance due to the great results that immunotherapy has achieved in some patients. Currently the microsatellite instability seems to be the best biomarker to select CCR patients who may benefit from this kind of treatment[1]. However only one third of these patients response to immunotherapies[17]. This situation makes necessary to develop a better stratification system to select patients for immunotherapy treatments, being microenvironment status one of the elements that could affect more. In this work we have studied the immune subtype classification in CRC in depth, obtaining results that suggest it possible relevance in the patients stratification based on in the tumor microenvironment and consequently, it's importance in the patient selection to immunotherapy.

INDEX

1. INTRODUCTION

1.1 Epidemiology and origin of cancer

Cancer has cost 9.6 million of lives in 2018, being a leading cause of death worldwide (https://www.who.int), in our country the number of cases estimated in 2017 was 228.482 but to 2.035 is expected around 315.400 new cases (https://seom.org/). This pathology involves a great number of diseases that could affect different tissues. The common pattern of these pathologies is the accumulation of alterations at genomic level that promote the uncontrolled and aggressive growth of the cells, originating a malign tumor. Those cells can acquire selective advantage, allowing its perpetuation and the development of invasive and migration skills which cause metastasis events[18].

The way to obtain the genomic alterations makes etiology of cancer is diverse. It can develop due to inheriting an alteration from one of the progenitors, what is known as hereditary cancer. Nevertheless, hereditary cancer involves only between 5-10% of patients, so most of the alterations are somatic (genomic alteration generated in non-germline cell), developing what it is called sporadic cancer[19]. Inside of sporadic cancer we can found several factors that influence the development of cancer. It could be endogenous (hormones and metabolites) or exogenous (alcohol, tobacco, drugs, radiation viruses)[20].

In the last decade, several international cancer genomes research consortiums have been established with the aim of comprehensively characterize the genomic alterations of several tumors. An example is The Cancer Genome Atlas (TCGA)

consortium (http://www.cancergenome.nih.gov/) which has characterised over 20,000 primary cancer from generating the biggest database with -*omics* cancer information. This meant over 2,5 petabytes of free information available for the research community divided into genomics, epigenomics, transcriptomics, proteomics and clinical data. TCGA also allowed the creation of PanCancer, an international research project called focused on the study of the differences among genomic and molecular alterations across diverse tumor types[21].

The information provided by the cancer genome sequencing projects have allowed to reveal 'driver genes', that is, those altered genes which confer a selective growth advantage to tumor cells[22]. It has been defined 2 types of driver genes [23].

- Proto-oncogenes: Genes that are related to survival and cell division. The alteration in these genes affect their own regulation, making them constitutively activated. Usually only one altered allele of these genes is needed to generate the carcinoma cell phenotype.
- Tumor suppressor genes: Genes that prevent the tumor development by modulating cell cycle progression and apoptosis. Alterations of these genes generate the loss of function of the protein, avoiding the correct regulation of cell growth. In this case both alleles must be altered to show the tumor phenotype.

1.2 Molecular basis of colorectal cancer

CRC is the neoplasia with more incidence in the world considered both genders. Furthermore, is the second dead cause in men and women separately. Nevertheless, the early detection using better screening protocols has entailed the 5-year survival improvement on CRC patients, but a 25% of patients still develop metastasis whose survival rate at 2 years around 35%[24].

As other types of cancer, the origin of CRC is associated with the mutation accumulation obtained by inheriting or by exogenic and endogenic factors. Thus, promote the passage from normal epithelium to invasive carcinoma[25]. One of the exogen factors more related to CRC development is the diet. Some eating habits that may influence to develop CRC cancer are the triglycerides and red meat excess consumption joined to low presence of fibber, D vitamin and phosphates on diet. Besides, some metabolites generated by bacterial flora have been related to the colorectal carcinogenesis[26,27]. The main driver genes in CRC are the proto-oncogenes KRAS and BRAF and the tumor suppressors APC, PTEN and TP53 which alteration causes the tumor progression[19,28]

CRC tumor can be classified into two main groups from a genomic perspective; hypermutated tumors and no hypermutated tumors. Three-quarters of hypermutated tumors have microsatellite instability (MSI), generally due to hypermethylation and silencing of the MLH1 promoter, and one-quarter have somatic mutations in mismatch-repair (MMR) or polymerase ε (POLE) genes. On the other hand, the majority of sporadic colorectal tumors (~85%) present chromosomal instability (CIN) and are generally non-hypermutated. The pattern of genomic and epigenomic events, thus, substantially differ in hypermutated versus non-hypermutated tumors[28].

1.3 Immunotherapy in CRC

Immunotherapy is defined as cancer therapies, which mechanism of action consists in stimulating immune response to tumors. There are different types of immunotherapies such as CART-t cell therapy, cytokine therapy or cancer vaccines, but the most used in cancer are immune checkpoint inhibitors[29]. Immune checkpoints are control proteins present on the surface of the cell. The union between this proteins and T cell ligands promote the inhibition of this lymphocytes. This mechanism is necessary in the body to avoid the development of autoimmune reactions but tumors use it to escape from the immune system. The use of molecules that block this unions have raised in the recent years because of the great observed benefit in some patients. However, the lack of response in others makes necessary the correct stratification of the patients to decide to treat them with immune checkpoints inhibitors[29]. In CRC, immunotherapy was approved in 2017 but only for heavily mutated tumors that are mismatch-repair-deficient (dMMR) or have high levels of microsatellite instability (MSI), since patients with tumors that do not accomplish these requirements show resistant to these drugs. However, the recent clinical trials establish that the response rate of the tumor with this characteristic is between 50 – 69%, depending on study[4].

1.4 Cancer transcriptomics

Transcriptomics technologies are the techniques used to study an organism's transcriptome, the sum of all of its RNA transcripts[30]. Currently, there are two main high-throughput technologies which are used to obtain transcriptomics data: i) gene expression microarrays and ii) RNA-seq. Gene expression microarrays were born in 1995, this technology is based on hybridization between the cDNA of a sample (RNA that has been

retro-transcribed to DNA) and fluorescent probes that, usually, represent the known transcriptome[30]. Gene expression microarrays allowed to quantify expression level of each gene (represented by probes) were being displayed in the sample, what meant a great step in the research of the biology of cancer. Some gene expression microarray limitations include the need of an *a priori* knowledge of the transcriptome sequence to design the probes, cross-hybridization events and need a high amount RNA[30].

RNA-seq consist on sequencing fragments of cDNA that calls reads. These reads are aligned with a transcriptome reference. Reads which belong to the same gene are quantified and normalized considered the gene size and the number of total reads into RPKM (reads per kilobase million) or FPKM (fragments per kilobase million). Finally, the reads per gene are obtained in all samples giving the gene level expression[30]. Differences between these two transcriptomics technologies have been discussed in other publications[30,31].

The most usual applications of transcriptomics high-throughput technologies are focused on comparing two conditions (e.g. tumor versus control), obtaining the differentially expressed genes between them [30], performing functional analysis[32] and establishing molecular classifications[6,8,15].

1.5 Transcriptomics based classification systems in CRC.

CRC in one of the tumors which more people have studied to make a classification system based on transcriptomics data, existing eight different classifications before 2015[9–14]. These different classification systems have some similarities, but the lack of consistency among them triggers the creation of the international Colorectal Cancer

Subtyping Consortium (CRCSC)[15]. The CRCSC developed the consensus molecular subtypes, based on the previous classification systems. Consensus molecular subtypes (CMS) classification distinguish among four different consensus molecular subtypes: i) CMS1 (MSI immune subtype, 14% of early-stage tumors), characterized by tumors hypermutated, hypermethylated, enriched in BRAFV600E mutations and with a strong immune activation; ii) CMS2 (canonical subtype, 37% of early-stage tumors), includes tumors with high chromosomal instability and somatic copy number alterations, with a marked upregulation of WNT and MYC signalling, and higher expression of epithelial features; iii) CMS3 (metabolic subtype, 13% of early-stage tumors), encompasses epithelial tumors with metabolic deregulation and enriched in KRAS-activating mutations and iv) CMS4 (mesenchymal subtype, 23% of early-stage tumors), defined by a strong activation of epithelial-mesenchymal transition (EMT), angiogenesis, transforming growth factors (such as TGF-β) and stemness pathways, that is the CMS with the worst relapse-free and overall survival rates[15].

The CMS classification is the most robust classification currently available for the stratification of CRC and has showed to have a prognostic value. On an immune point of view, CMS1 is the subtype with more abundance of cytotoxic T cells (CD8+) and macrophages (CD68) cells followed by CMS4, which also is characterised by having a high level of stromal and fibroblast cells. On the other hand, CMS2 and CMS3 subtypes are cold tumors with less presence of immune populations. CMS1 also present more expression of chemokine coding genes related to T cell activation (CXCL9, CXCL10, CXCL16, IFNG and IL15) and genes involved in T cell inhibition (PDCD1, CD274, PDCD1LG2, CTLA4 and LAG3). However, tumors which belong to CMS4 subtype have more activated genes with important role in angiogenesis(VGFB, VEGFC, PDGFC), genes related with immunosuppression (CXCL12 , TGFB1, TGFB3 and

LGALS1) and genes expressed in complement system cells (C15R, C1R, C3, C3AR1, C5AR1, CT, CFO, CFH and CF1)[33].

1.6 Immune subtype (IS) classification

The relationships between the immune system and cancer have acquired an important impact in the recent years. Currently there are several researches that establish it clear prognostic value and the effect to the response of several anti-cancer therapies[34]. With this motivation Thorsson *et al* have developed an IS classification in PanCancer[16]. This IS classification is made over more than 11.000 patients from all 33 non-haematological TCGA cancer types. They used a scored collection of 160 different immune expression signatures with single sample gene set enrichment (ssGSEA) to identify modules of immune related gene sets. Then they chose five signatures where each one represents a module and they made a clustering using these scores. As a result, they identified 6 distinct immune subtypes (C1-C6) described below, characterized by relevant differences in lymphocyte or macrophage signatures, T helper type 1 (Th1) : T helper type 2 (Th2) ratio, expression of immunomodulatory genes, intratumoral heterogeneity, neoantigen load, aneuploidy, cell proliferation and prognosis[16]:

- **Wound Healing subtype (C1)** showing an elevated expression of angiogenic genes, a high proliferation rate and a low Th1:Th2 ratio related to the adaptive immune infiltrate.
- **IFN-γ Dominant subtype (C2)** presenting a high proliferation rate too, the highest intratumoral heterogeneity, the highest Macrophage type 1 (M1):

Macrophage type 2 (M2) polarization and CD8+ T cell population, and the greatest TCR diversity.

- **Inflammatory subtype (C3)** defined by elevated Th17 and Th1 genes, low to moderate proliferation, lower levels of aneuploidy, higher somatic copy number alterations and the most favourable prognosis.
- **Lymphocyte Depleted subtype (C4)** presenting moderate cell proliferation and intratumoral heterogeneity, and a prominent macrophage signature with Th1 suppressed and a high M2 response; consistent with these features, it was associated with a poor outcome.
- **Immunologically Quiet subtype (C5**) displays the lowest lymphocyte and highest macrophage responses, dominated by M2 macrophages, and had low rates of proliferation and heterogeneity.
- **TGF-β Dominant subtype (C6)** was a small group of mixed tumors with the highest TGF-β signature and a high lymphocytic infiltrate with a balanced Th1:Th2 ratio. Together with C4, C6 was associated with the worst prognosis.

2. OBJECTIVES

The main goal of this work is to study the clinical value of combining IS and CMS classifications in CRC (Figure 1). To achieve this goal, we have used CRC TCGA data to study the clinical, molecular and immune differences among immune subtypes in the whole cohort and using each CMS cohort separately.

1. **To study the relationship between CMS and Immune subtype classification in CRC.** Proportion of Immune subtypes in each CMS and vice versa have been established to look for an enrichment in any of the two directions.
2. **To analyse the association between clinic-pathological variables and Immune subtypes in CRC.** Clinic-pathological variables have been studied to look for associations with any of the Immune subtypes in the general cohort and in each CMS.
3. **To observe differences in the proportion of distinct immune populations among immune subtypes.** We have analysed differences in the abundances of the distinct immune population cells, using the whole cohort and in each CMS.
4. **To obtain the differentially expressed genes and the enriched pathways among the two principal immune subtypes in CRC.** We have assessed the genes and biological pathways in which the two main immune subtypes are different. We are going to use all patients together and distinguish them by CMS.
5. **To investigate the prognostic and potential predictive power of immune classification in CRC.** We have analysed overall survival data to study possible prognostic value of Immune subtypes in CRC. Besides, we have examined the expression of important immunomodulator and immunotherapy biomarker genes.

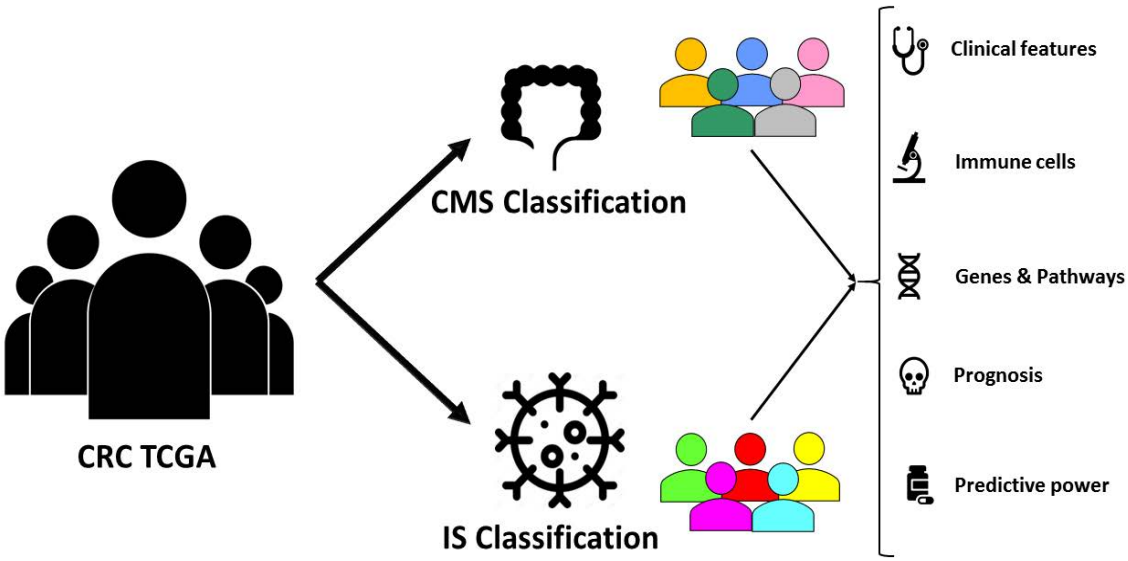


Figure 1: Summary of the main objectives and steps encompassed in this project. CMS and IS classifications are integrated in CRC patients in order to study the overlapping between them. Furthermore, differences in principal clinical variables, immune population abundances, genes expression, prognosis and potential predictive power are assessed among immune subtypes in the whole cohort and in each CMS.

3. MATERIAL AND METHODS

3.1 Study population and datasets

3.1.1. Study population

The study population included all colorectal cancer patients from TCGA database (N= 625)[35]. We have selected those patients with available information on the IS as per the new immune classification recently developed by Thorsson et al (N=597)[16]. Information about IS were obtained from the Immune Subtype Clustering github repository (https://github.com/Gibbsdavidl/Immune-Subtype-Clustering/tree/master/shiny-app/Immune-Subtype-Clustering). Information on microsatellite instability, CIMP status and CMSs (N=573) was downloaded from the Colorectal Cancer Subtyping Consortium (CRCSC) Synapse (https://www.synapse.org/#!Synapse:syn2623706/wiki/67246)[15]. (Figure 2)

To obtain clinical and genomic data of TCGA patients, different data portals have been used. Clinical and RNA-seq data was acquired from the GDC colon and rectal Xena Browser data set (https://xenabrowser.net)[36]. TCGA data was also downloaded from cBioportal (http://www.cbioportal.org/datasets) where CNV data (N=614) and mutation annotation format (MAF) files with gene mutation status (N=526) were obtained[37].

3.1.2. Datasets

Different data from CRC cohort of TCGA project was used for this work such as:

- **Clinical data:** Matrix where patients are rows and clinical information are columns. Clinical data used for each analysis are detailed in Supplementary Tables 1 and 2.

- **RNA-seq data:** Matrix with log2 transformed read counts per gene. Genes are in rows while samples are in columns. Read counts belong to TCGA version 3 data and is obtained by Htseq software[38]. Some data processing was made before using the datamatrix:

    1. Log2 transformation was reverted from data prior to data analysis.

    2. Samples from metastasis were deleted from expression matrix.

    3. Twenty five repeated genes were collapsed using the row with more variance among all samples.

- **CNV data:** Data generated by GISTIC software obtained[39]. CNV data was analysed considering only deep amplifications for oncogenes (log2CopyNumber>1) and deep deletions for tumor suppressor genes (log2CopyNumber<1).

- **Mutation data:** Matrix where each row represents one mutation with further information such as sample, gene, mutated allele, reference allele, chromosome position etc. Synonym mutations, and non-coding exon and intron mutations were filtered out from the MAF file prior to mutation detection and hypermutation class identification. Hypermutation phenotype threshold was established using a mutation rate>12 per 10^6 bps as previously described[35]. Then a matrix with

samples as rows and genes of interest as columns was made with mutation information where 1 is mutated and 0 is non mutated. It is important to highlight that only V600E mutation was considered for the analysis of BRAF mutations. The mutated genes analysed appear in Supplementary Tables 1 and 2.

3.2 Clinical and molecular data analyses

Descriptive statistics were used to characterize the most relevant clinical parameters. Fisher's exact test was applied to obtain the associations of categorical variables (clinical features and mutations) and colorectal cancer subtypes by using R software[40]. *P-values* were obtained by Monte Carlo simulation with 10^4 replicates.

3.3 Overall Survival

Overall Survival (OS) is the elapsed time from the date of diagnosis to the date of death of any cause or patient's last clinical consultation in alive patients. Survival function was estimated according to the *Kaplan-Meier* method[41] while differences in survival distributions between groups were assessed using the *log-rank test*[42] (Figure2). Cox proportional hazard univariate and multivariate analysis were also conducted, including relevant known clinical and pathological prognostic factors (age, gender, site of primary tumor, tumor histology and stage at diagnosis) as well as the IS and CMS classifications. Finally, the Cox proportional risk regression model was fitted to data to estimate the independent prognostic value in terms of OS and confound variables were analysed. The basic assumptions of the model were evaluated employing proportional hazards.

3.4 Immune cell population analysis.

Immune cell abundance estimation was generated applying deconvolution-based software to RNA-seq data. These programs use a precomputed matrix with information about transcriptomic markers of different immune cell population and extract this signal from a mixture RNA-seq data from bulk tissue (Figure2). The deconvolution software employed were CIBERSORT[43] and MCP counter[33]:

- CIBERSORT: CIBERSORT is a deconvolution method that estimate relative proportion of 22 cell type, using a leukocyte signature matrix and a linear support vector regression approach. Relative proportion of the 22 immune cell types analysed by CIBERSORT were obtained from supplemental material of The Immune Landscape of Cancer publication. Relative percentages were then multiplied by Leukocyte fraction to obtain Absolute fraction[16]. Statistical analysis to assess differences in the overall fraction of immune cell types between ISs were performed applying pair-wise Wilcoxon Rank test using R software[40].
- MCP counter: MCP counter is a deconvolution method which calculates the absolute abundance of 11 cell population, using specific transcriptomic markers. Immune and stromal cell populations were estimated by MCP-counter[33], using the upper quartile normalized and log2+1 transformed RNA-seq data. Statistical differences between ISs were evaluated using pair-wise Student's *t test* in R software[40].

Data related to tumor mutation burden (TMB), B cell receptor richness (BCR) and

T cell receptor richness (TCR) were consulted in supplemental material of The Immune Landscape of Cancer publication and differences among immune subtypes were applied using Wilcoxon rank test in R software[40]. ComplexHeatmap R package[44] and z-score values obtained from all samples were employed to generate immune population and IS heatmaps.

3.4 Differential expression and gene set enrichment analysis

Differential expression analysis between wound healing and IFN-γ ISs was performed using Bioconductor's DESeq2 package[45]. To account for multiple hypotheses testing, the estimated significance level (P-value) was adjusted using Benjamini & Hochberg False Discovery Rate (FDR) correction[46]. Genes with FDR <0.05 were selected as differentially expressed between both subtypes. Prior to differential expression analysis, genes showing flat patterns (inter quartile range<0.5) across patients were filtered out from DESeq2 normalized matrix[48] (Figure2). CMS subtype and batch number available in clinical data were used as covariates in differential expression analysis in order to avoid potential confusion effects in analysis involving all patients. Gene annotations were obtained by biomaRt R package[47] and KEGG pathways were obtained from KEGG.db package[48] . Coincident genes found in all analysis performed were retrieved using UpSetR package[49] .

Gene set enrichment analysis was computed with ssGSEA algorithm implemented in Bioconductor's gsva R package[50] , using upper quartile normalized RNA-seq data and gene sets of Hallmarks and KEGG collections available in MSigDB (http://software.broadinstitute.org/gsea/msigdb/index.jsp). Differentially activated

pathways between the two predominant ISs, Wound Healing and IFN-γ Dominant sub types, were assessed performing the Student's T test implemented in RVAideMemoire[51] with 1000 permutations to the gene sets enrichment scores (ES). The resulting p-values were adjusted for multiple comparisons by Benjamini-Hochberg's FDR correction[46] (Figure2). ES median Z-scores from significant differentially enriched pathways among Wound Healing and IFN-γ Dominant subtypes were plotted in a heatmap using ComplexHeatmap R package[44].

3.5 Analysis of immune modulator genes

Gene expression analysis: only those samples with RNA-seq data available were employed for immune modulator genes expression analysis (N=590). To this end, we applied upper quartile normalization and log2(expression+1) transformation for each of the 75 immune modulator genes with respect to each immune subtype. Then we performed a Kruskal-Wallis or Wilcoxon signed-rank test depending on the number of immune subtypes considered in each analysis. The resulting p-values were adjusted for multiple comparisons by Benjamini-Hochberg's FDR correction. ComplexHeatmap R package was employed to build heatmaps using Z-score median values from all immune modulator genes[44].

CNV analysis: only samples with CNV data were included in this analysis (N=432). Immune modulators CNV analysis was carried out considering deep amplifications (log2CopyNumber>1), shallow amplifications (log2CopyNumber=1), shallow deletions (log2CopyNumber=-1) and deep deletions (log2CopyNumber<-1). We calculated the differences between observed and expected frequencies for every immune

modulator gene in each immune subtype. Heatmaps were obtained using ComplexHeatmap R package[44].

3.6 Supplementary material

Supplementary figures, tables and results are available at http://bit.ly/2Zz3Frh

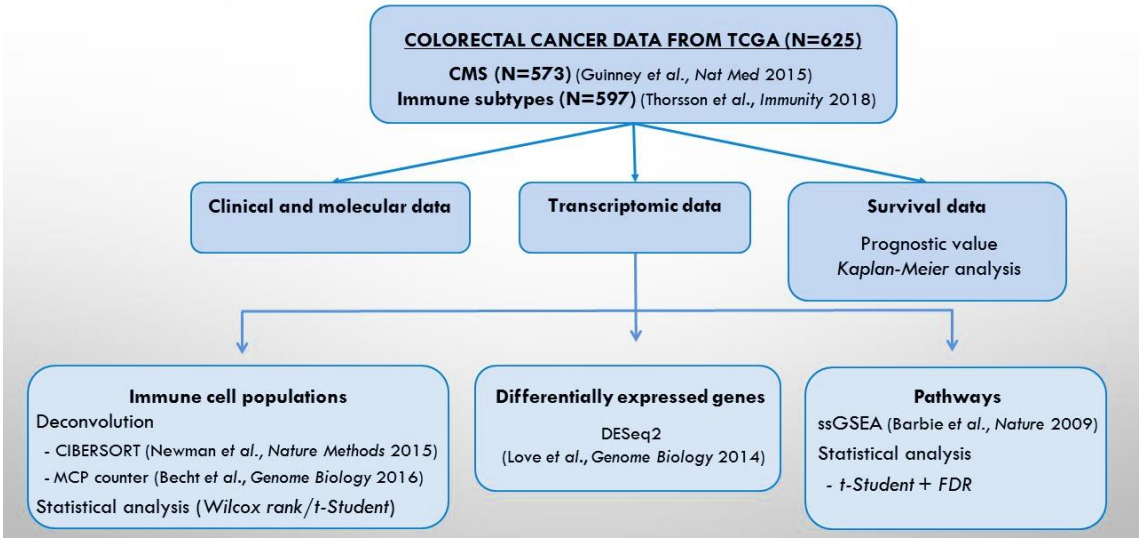


Figure 2: Resume of the material and methods used in this project. Consensus Molecular Subtypes and Immune subtypes were obtained from TCGA patients cohort. Clinical, survival and transcriptomics data of these patients were analysed. Transcriptomics data was used to obtain immune cell population applying deconvolution methods (CIBERSORT and MCP counter), differentially expressed genes (DESeq2) and differentially enriched pathways (ssGSEA + *t-Student* test).

4. RESULTS

4.1 Overview

In this project we perform the integration of CMS and IS transcriptomics classification system (Figure 3). Such analysis allow us to obtain the following results: 1) the Immune Subtype distribution in the global CRC population and by CMS (see section 4.2), 2) clinical and pathological features of colorectal cancer patients by immune subtype (section 4.3), 3) the prognostic impact of immune subtypes in colorectal cancer (section 4.4), 4) immune and stromal cell population in distinct immune subtypes of CRC patients (section 4.5), 5)expression of immunomodulatory genes (section 4.6) 6) genes and pathways differentially expressed among the principal immune subtypes (section 4.7).

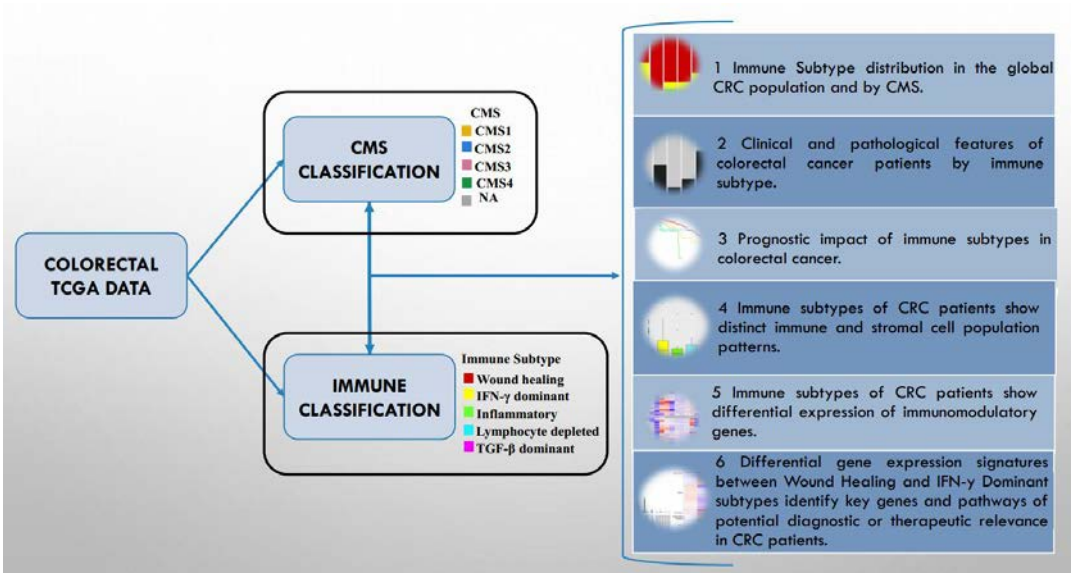


Figure 3: Project and results overview.

4.2 Immune Subtype distribution in the global CRC population and by CMS

First, we assessed the distribution of immune subtypes in the global CRC population (N=573). We found that the 2 predominant ISs in CRC were the C1 Wound Healing subtype, identified in 459 samples (77%), and the C2 IFN-γ Dominant subtype, present in 103 samples (17%). The rest of ISs were barely present in CRC patient as we saw in C3 Inflammatory subtype (18 samples, 3%), C4 Lymphocyte Depleted subtype (13 samples, 2%) and C6 TGF-β Dominant subtype (4 samples, 0.7%) subtypes. Of note, no CRC sample belonged to the C5 Immunologically Quiet IS (Supplementary Table 1).

Then, we evaluated the crossover between CMS and IS classifications (Figure 4, Supplementary Tables 3 and 4). The 2 main CRC ISs, C1 and C2, were present in all CMSs but their relative distribution is not equal in all CMSs (Figure 4A, Supplementary Table 3). In fact, the C1 Wound Healing IS was markedly dominant in CMS2 (91%) but far less common in CMS1 (46%) tumors, whereas the proportion of C1 Wound Healing in CMS3 and CMS4 tumors was similar to the average observed for the overall CRC population (77-78%). On the contrary, the C2 IFN-γ Dominant was the most common IS observed in CMS1 tumors (53%) and was under-represented in CMS2 tumors (8%), whereas the proportion of the C2 IFN-γ Dominant in CMS3 and CMS4 tumors was slightly higher (11-13%). Other ISs were barely present in CMS1 and CMS2 tumors (Lymphocyte Depletion subtype – 1 patient in each CMS group; Inflammatory subtype – not represented in the CMS1 group, 2 patients in the CMS2 group; no TGF-β Dominant cases). The immunological landscape of CMS3 and CMS4 was slightly more diverse: they have more patients from C3 Inflammatory subtype (7% and 6%, respectively), as

compared to CMS1/2. Besides, CMS3 had the highest representation of the C4 Lymphocyte Depleted subtype (4%), and all 3 cases with the TGF-β Dominant phenotype belonged exclusively to the CMS4 subgroup (2.3%). Then, we analysed the distribution of CMS groups by immune subtype (Figure 4B, Supplementary Table 4). We found that the C1 Wound Healing subtype is enriched in CMS2 and CMS4 groups (46% and 25%, respectively) and the remaining classified samples were distributed between CMS1 (8%) and CMS3 (13%). Eight percent of the C1 Wound Healing samples were not classifiable in any CMS group. The C2 INF-γ Dominant samples are enriched in CMS1 (41%), although a notable proportion of patients belonged to other CMS groups (17% were CMS2 and 18% CMS4), and a small fraction was classified as CMS3 (9%). The distribution of CMS in other less common IS in CRC is less reliable due to the limited number of samples within each category. However, a relevant proportion of Inflammatory and Lymphocyte Depleted samples were classified as CMS3 (28% in both cases) and CMS4 subtypes (44% and 18%, respectively). The remaining samples of the Inflammatory subtype were classified as CMS2 (11%), while the rest of Lymphocyte Depleted samples were equally distributed in the CMS1 and CMS2 groups (9% each). Finally, all 3 TGF-β Dominant samples were classified as CMS4 (100%).

In summary, these results indicate that the subtypes most frequently found in CRC were Wound Healing and INF-γ Dominant subtypes while Inflammatory, Lymphocyte Depleted and TGF-β Dominant are less represented. Furthermore, not all CMS has the same proportions of IS, being CMS1 the most differentiated with a high proportion of INF-γ Dominant samples.

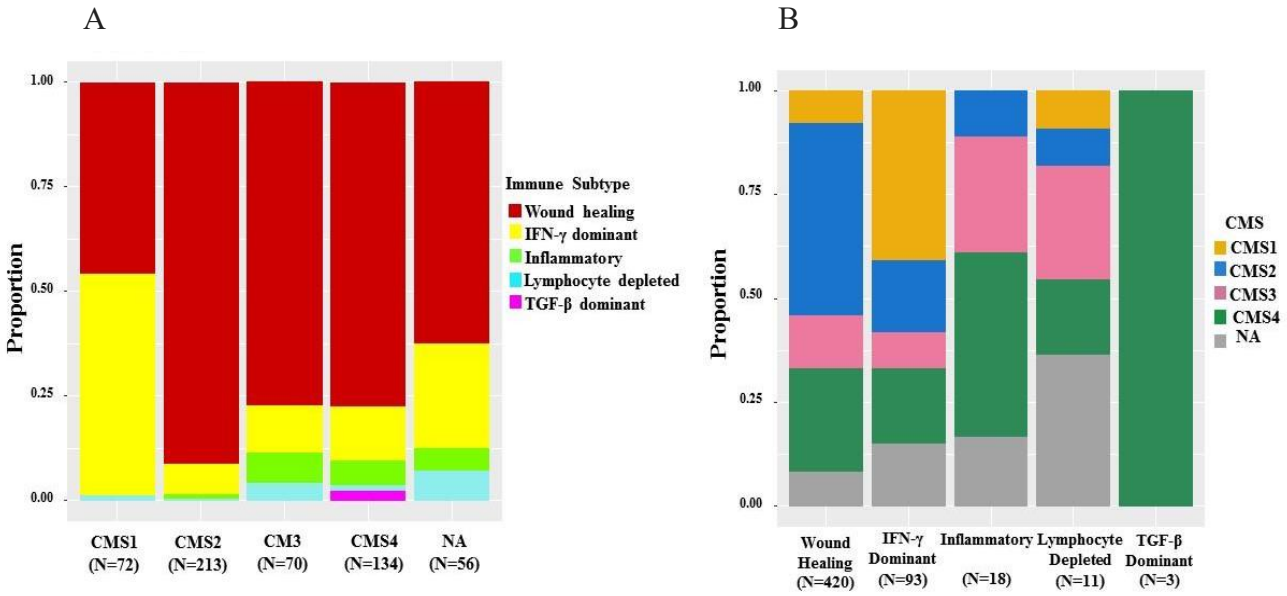


Figure 4: Interplay between Immune subtypes and CMS in CRC: A) Distribution of ISs in CMS1, CMS 2 CMS3 and CMS4 subtypes. B) Distribution of CMSs in Wound Healing, IFN-γ dominant, Inflammatory and TGF-β dominant subtypes.

4.3 Clinical and pathological features of colorectal cancer patients by immune subtype

Main clinical, pathological and molecular characteristics by immune subtype in the global CRC study population are depicted in Figure 5 and detailed in Supplementary Table 4. In summary, we found that C2 IFN-γ Dominant is enriched in tumors located in the right colon, while C1 Wound Healing and C3 Inflammatory ISs were frequently encountered in left-sided tumors. No significant association with ISs were observed by gender or stage at diagnosis. All immune subtypes presented a similar proportion of KRAS mutations, with the exception of the TGF-β Dominant IS, were no KRAS mutations were identified. On the contrary, enrichment of BRAF mutations was observed in the IFN-γ Dominant subtype (25%), and also in the Lymphocyte Depleted (17%) and TGF-β Dominant (33%) subtypes, although these figures are not too reliable due to the small sample size in these last two subgroups. MSI, CIMP and hypermutated phenotypes

were consistently more frequently found in the IFN-γ Dominant subtype. The incidence of APC mutations was highest in C1 Wound Healing, and TP53 mutations observed in >50% of C1 Wound Healing or C2 INF- γ Dominant ISs but were less common y C3 Inflammatory and C4 Lymphocyte Depleted ISs. However, we repeated this analysis using only patients that belong to the same CMS group and no significant relationship were found in any of the previous associated variables (Supplementary Figures 2-5 and Supplementary Tables 5-8). The incidence by immune subtype of the 10 most frequently genetic alterations in hypermutated and non-hypermutated tumors were provided in Supplementary Table 9. Connections observed between different clinical, pathological and molecular characteristics and CMSs are consistent with those previously reported (Guinney *et al.,* 2015) (Supplementary Figure 1 and Supplementary Table 3).

These results suggest that the associations found between clinical features and ISs are brought by the different proportion in CMSs, but not by itself.

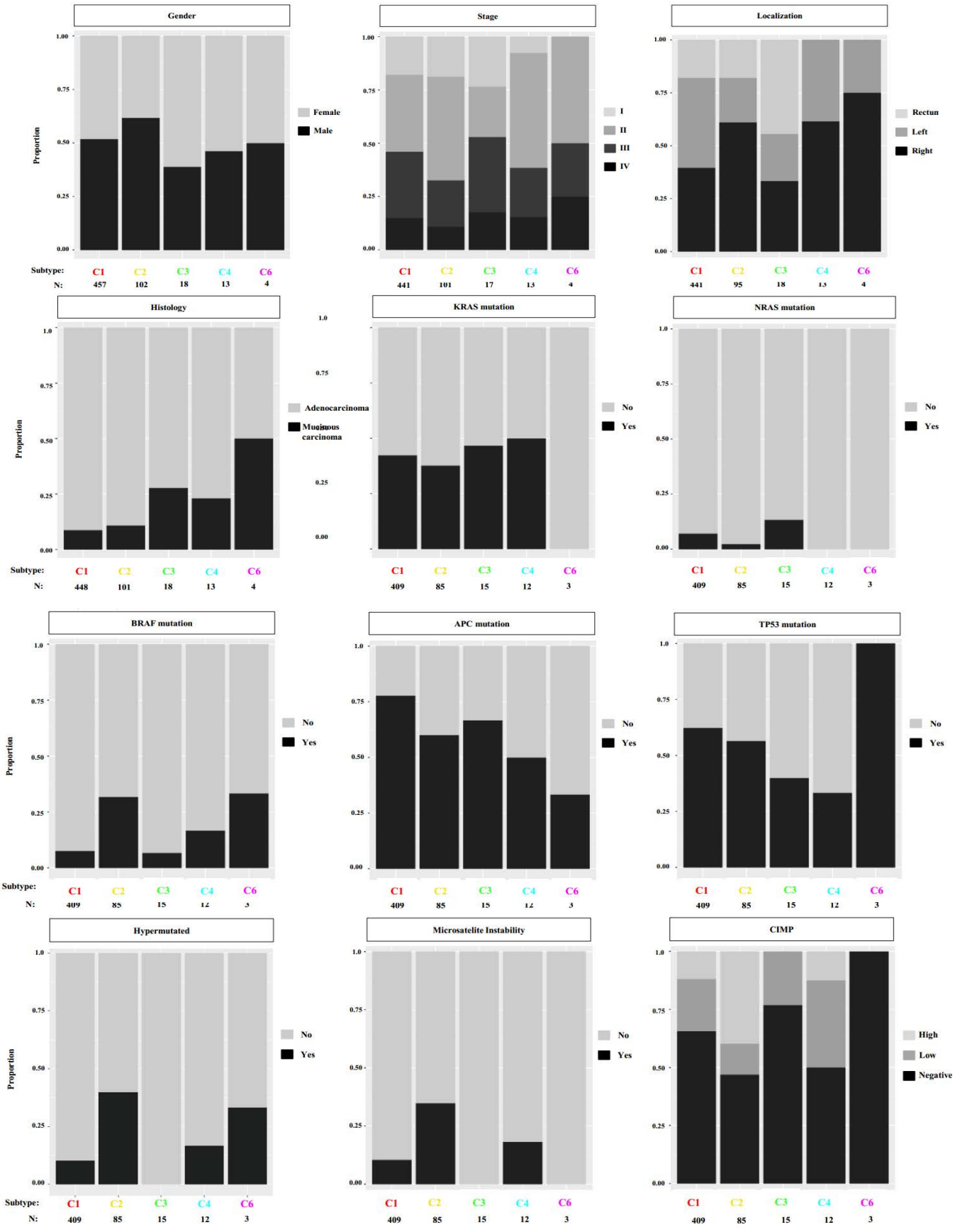

Gender
Female
Male
Stage
I
II
III
IV
Localization
Rectun
Left
Right
Histology
Adenocarcinoma
Mucinous carcinoma
KRAS mutation
NRAS mutation
BRAF mutation
APC mutation
TP53 mutation
Hypermutated
Microsatelite Instability
CIMP
High
Low
Negative
No
Yes
Proportion
Subtype:
N:
C1
C2
C3
C4
C6

Figure 5: Clinicopathological characteristics of colorectal cancer patients according to immune subtypes. Bar plots showing the proportion of gender, tumor stage, localization of primary tumor, histology, KRAS, NRAS, BRAF, APC and TP53 mutations, hypermutated phenotype, microsatellite instability and CpG island methylator phenotype (CIMP) in immune subtypes C1 (wound healing), C2 (IFN-γ dominant), C3 (inflammatory), C4 (lymphocyte depleted) and C6 (TGF-β dominant). Number of patients with specific clinical information for each immune subtype is detailed below the bar plots.

4.4 Prognostic impact of immune subtypes in colorectal cancer

The OS analysis in CRC patients showed that the impact of the different ISs in survival significantly (*P= 0.02*) differed from that reported for the overall TCGA dataset, which included 33 different solid tumor types[16] (Figure 6A). Indeed, in the CRC study cohort, patients with the TGF-β Dominant and Wound Healing ISs showed better prognosis than other subtypes (5-year OS rates of 100% and 65%, respectively), whereas IFN-γ Dominant and Inflammatory were the subtypes associated with the worse outcome (5-year OS: 49% and 23%, respectively) (C2 vs C1, *HR 1.59, P=0.004*; C3 vs C1, *HR 2.77, P=0.02*) (Figure 6B, Supplementary Table 10). However, the lack of patients in some of the ISs makes impossible to draw practical conclusions. Similar trends were observed when we analysed the prognostic impact of immune subtypes within each CMS subgroup (Supplementary Figure 6 and Supplementary Table 10). Cox multivariate regression analysis showed that the immune subtype classification was significantly associated with survival in CRC patients, independent of other well established prognostic factors such as age, stage at diagnosis, or localization of primary tumor (Supplementary Table 11), showing that the immune phenotype had a significantly greater influence on OS than CMS (not significant in the multivariate analysis).

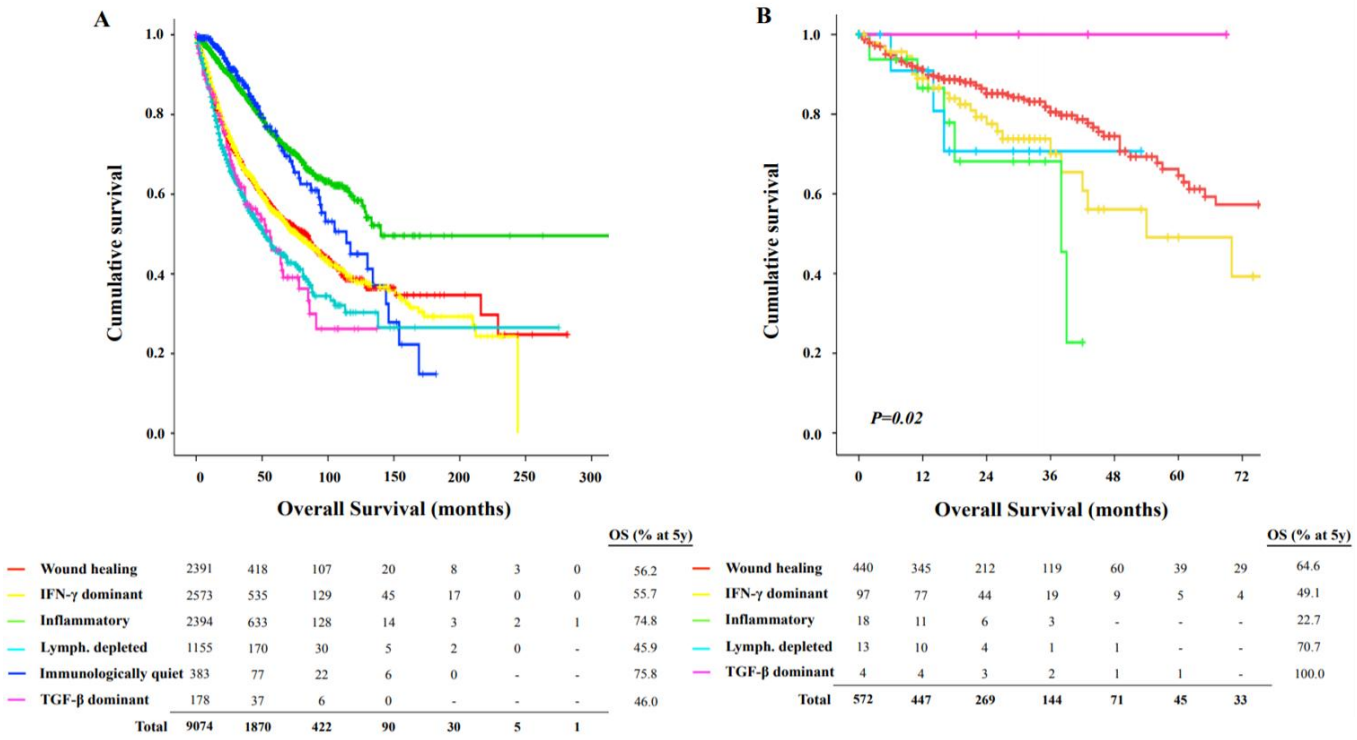
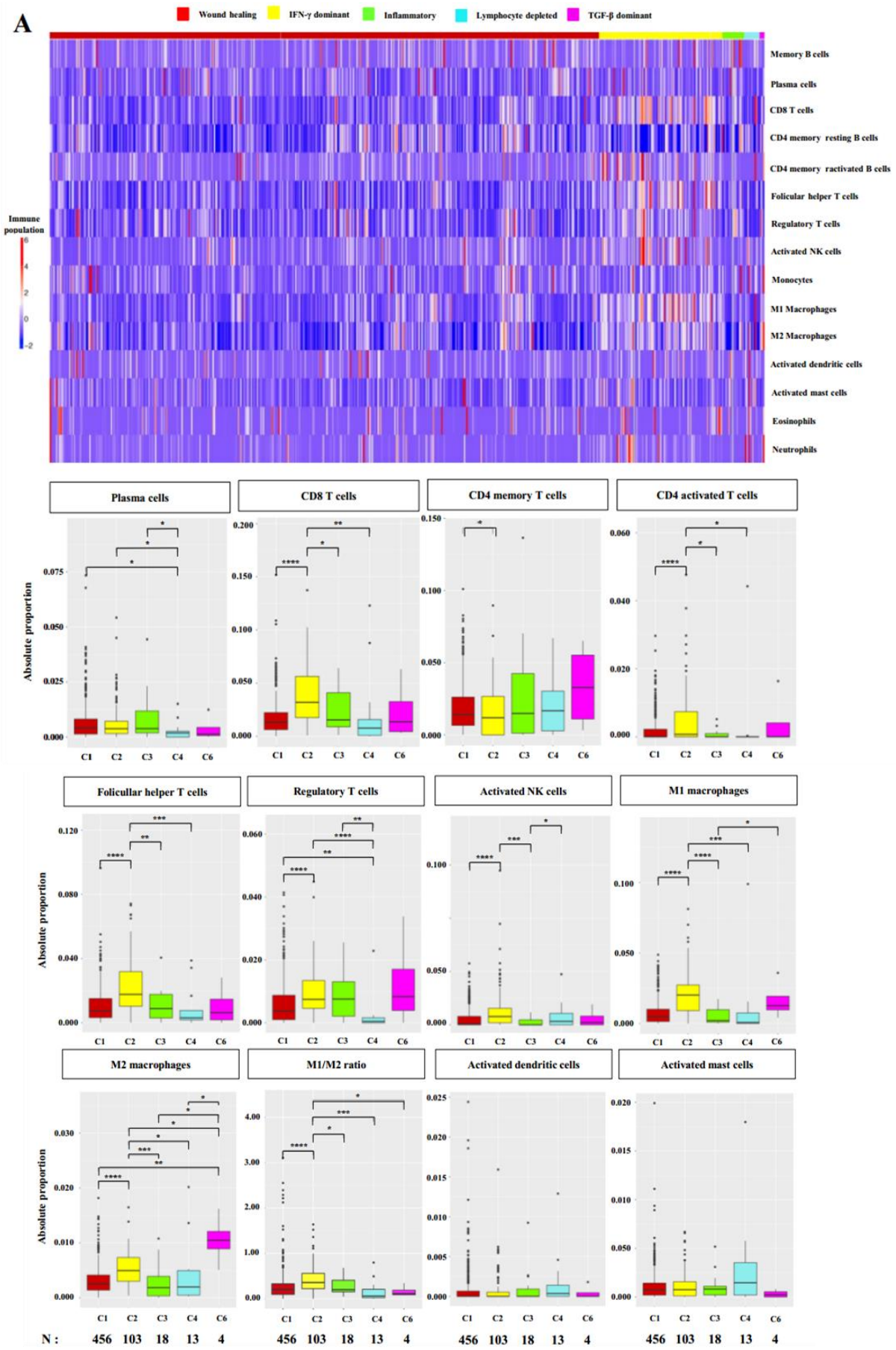


Figure 6: Impact of immune subtypes on survival of colorectal cancer patients. Patients at risk at the corresponding time point and survival rates at five years in each subtype are provided. P-value was calculated by log-rank test. CMS: Consensus Molecular Subtypes.

4.5 Immune subtypes of CRC patients show distinct immune and stromal cell population patterns

Different patterns for immune and stromal cell populations abundance were found among immune subtypes in CRC patients. The most relevant differences were identified between the two major subgroups, Wound Healing (N= 452) and IFN-γ Dominant (N=103). Thus, tumors with IFN-γ Dominant profile presented a strong expression of CD8+, CD4+ activated, follicular helper T cells, regulatory T cells, memory B cells and M1 macrophages. Moreover, high levels of natural killer cells (NKs), dendritic cells, neutrophils and M2 macrophages were also found in this subtype (Figure 7A and B). In contrast, Wound Healing subtype tumors presented higher proportions of memory CD4+

(Figure 7A). No differences were found in endothelial cells and fibroblasts abundance between these two groups, although the proportion of these populations was high in the four samples belong to TGF-β subtype (Figure 7B). Next, we analysed the distribution of immune and stromal cell populations by immune subtype within each CMS group. The distribution of all cell types analysed was similar in the CMS1 group, with significant differences observed between the Wound Healing and IFN-γ Dominant subtypes, except neutrophils and dendritic cells which do not have significant differences. (Supplementary Figure 7). In CMS2 immune and stromal cell proportions variate less between these two subtypes, being even more diluted in CMS3 and CMS4 (Supplementary Figures 8-10). These results show that IFN-γ is the subtype with more immune cell abundances which are also maintained in CMS1.


A
Wound healing
IFN-γ dominant
Inflammatory
Lymphocyte depleted
TGF-β dominant
Immune population
Memory B cells
Plasma cells
CD8 T cells
CD4 memory resting B cells
CD4 memory ractivated B cells
Folicular helper T cells
Regulatory T cells
Activated NK cells
Monocytes
M1 Macrophages
M2 Macrophages
Activated dendritic cells
Activated mast cells
Eosinophils
Neutrophils
Plasma cells
CD8 T cells
CD4 memory T cells
CD4 activated T cells
Folicullar helper T cells
Regulatory T cells
Activated NK cells
M1 macrophages
M2 macrophages
M1/M2 ratio
Activated dendritic cells
Activated mast cells
Absolute proportion
C1
C2
C3
C4
C6
N : 456 103 18 13 4

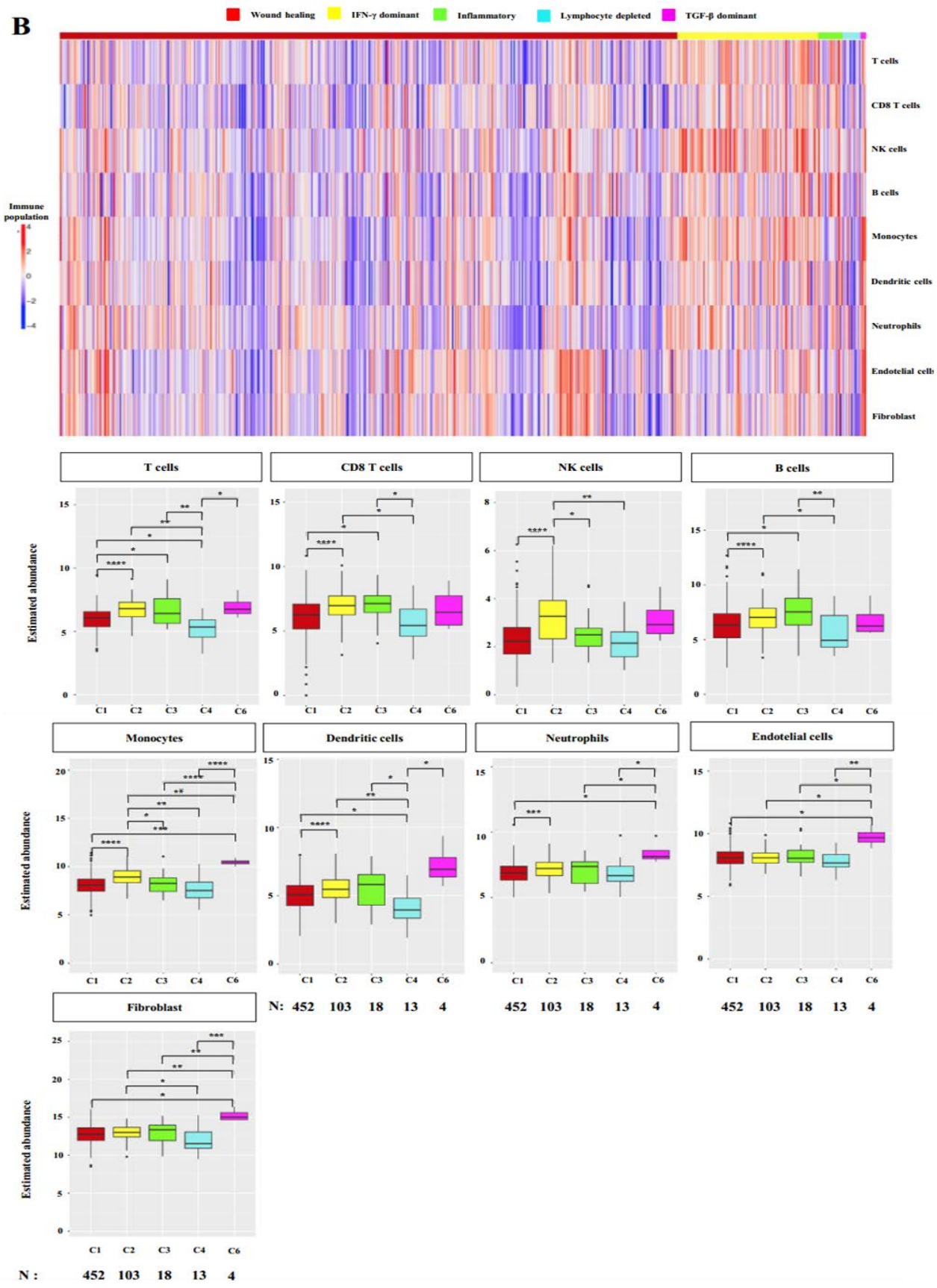

B
Wound healing
IFN-γ dominant
Inflammatory
Lymphocyte depleted
TGF-β dominant
Immune population
T cells
CD8 T cells
NK cells
B cells
Monocytes
Dendritic cells
Neutrophils
Endotelial cells
Fibroblast
Estimated abundance
C1
C2
C3
C4
C6
N: 452 103 18 13 4

Figure 7: Immune and stromal differential signatures among immune subtypes. A) CIBERSORT heatmap showing the distribution of lymphoid and myeloid lineages by immune subtype. *P<0.05, **P< 0.01, ***P<0.001, **** P<0.0001 using Wilcoxon test. B) MCPcounter heatmap showing the estimated abundance of several immune and stromal-cell populations in CMS1 colorectal cancer patients by immune subtypes. *P<0.05, **P< 0.01, ***P<0.001, **** P<0.0001 using Student's t-test.

4.6 Immune subtypes of CRC patients show differential expression of immunomodulatory genes

We examined gene expression, amplification or deletion of a large set of immunomodulators (IMs) across immune subtypes. Upregulation of several genes with different immunomodulatory functions were found in the IFN-γ Dominant subtype, such as CD80 (*FDR=1.69E-12),* CD28 (*FDR=5.20E-09)*, CD274 (PDL-1) (*FDR=6.80E23)*, CXCL10 (*FDR=1.83E-24)*, LAG3 (*FDR=6.91E-21)*, ICAM1 (*FDR=1.41E-11)*, HLA-DQA1 (*FDR=3.84E-16)*, HLA-DRA (*FDR=3.38E-17)*, IDO1 (*FDR=8.44E-20)*, and GZMA (*FDR=3.34E-18)* (Figure 8A and Supplementary Table 12). More specifically, immune inhibitors with the greatest differences between subtypes included PDL-1, PD1, CTLA-4, IDO1, or LAG3, and were most highly expressed in the IFN-γ Dominant subtype (Figure 8B). Significant differences observed between the 2 major immune phenotypes, Wound Healing and IFN-γ Dominant subtypes, were maintained within each CMS subgroup (Supplementary Figure 11).

Copy-number variations affected multiple IMs and both amplifications and deletions were frequently found in the Wound Healing subtype, while other immune subtypes showed fewer alterations (Figure 8A). Thus, SLAMF4, TNFSF4, CX3CL1, IL10, IL2, ENDRB, CD40, IDO1 and HMG β1 genes were most frequently amplified,

while VTCN1, CD276 (PDL-1), IL4, IL13, TNGRS-F18, -F4,-F9, -14 and ADORA2A were most frequently deleted. Overall, these observations are very relevant for cancer immunotherapy as numerous IM agonists and antagonists are currently under clinical development[29] . Moreover, there were significant differences in tumor mutational burden (TMB) which is larger in IFN-γ Dominant subtype. However, when we look at CMS1 cohort not differences were found between Wound Healing and in IFN-γ Dominant subtype. Related to B cell receptors (BCR), Lymphocyte Depleted subtype has less richness that the others. The same happens with T cell receptors (TCR) which are more richness in Inflammatory subtype.

In summary immunomodulatory genes are more expressed in IFN-γ subtype than in Wound Healing subtype while the opposite occurs with the alterations in these genes. However, IFN-γ is the subtype with more TMB, being not maintaining when we analysed each CMS.

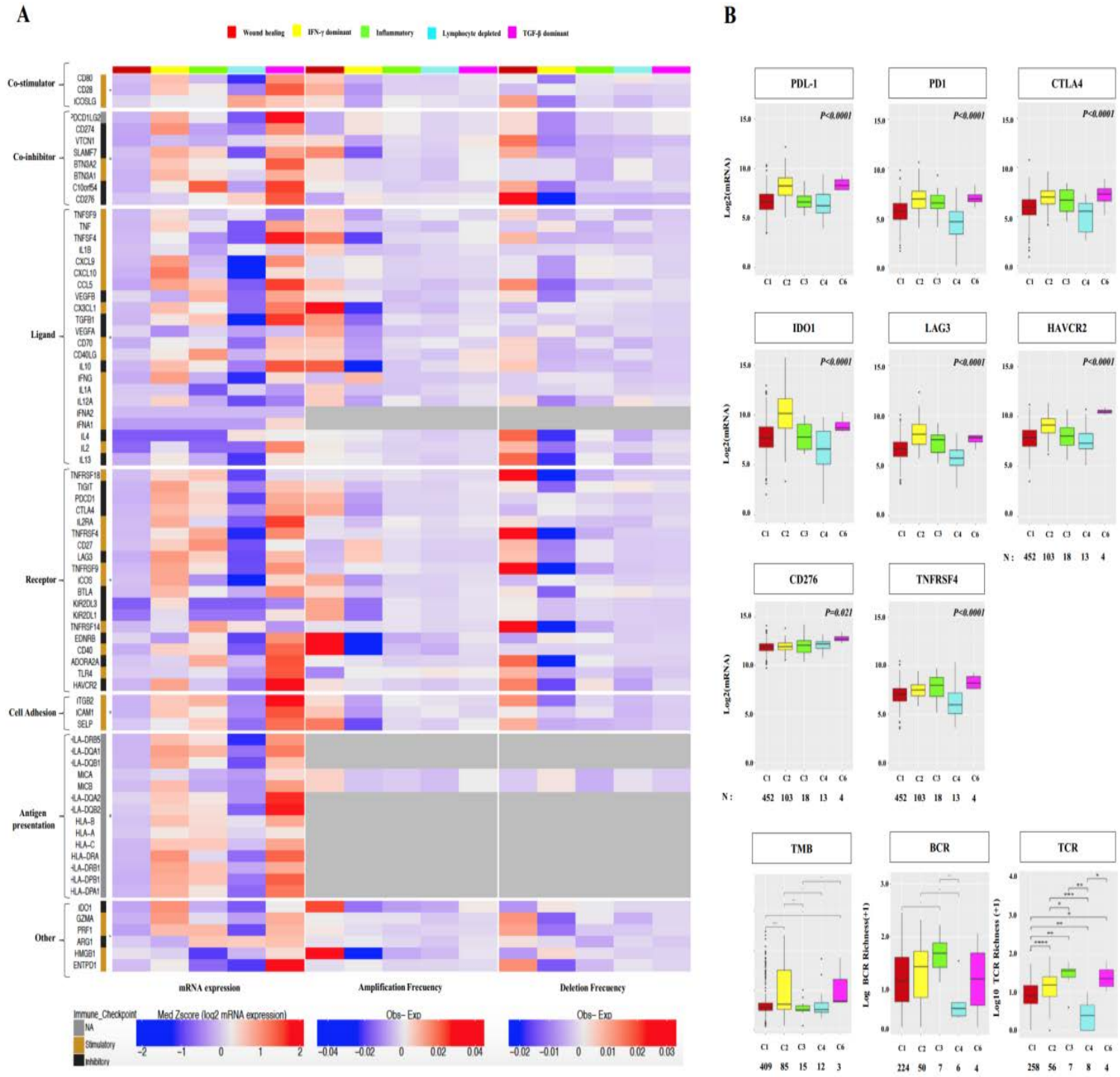


Figure 8: Expression and CNV of immune modulators (IM), tumor mutational burden (TMB), B cell receptor (BCR) and T cell receptor (TCR) diversity by immune subtypes. A) Heatmap showing median of log 2 normalized mRNA Z-score expression, amplification frequency (the difference between the fraction of samples in which an IM is amplified in a particular subtype and the amplification fraction in all samples); and the deletion frequency (as amplifications) for 75 IM genes by immune subtype. B) Expression levels of immune-

checkpoints; p value was computed using Kruskal-Wallis test. TMB, BCR Richness and TCR Richness were represented by immune subtypes and p values were computed by Wilcox rank test (*P<0.05, **P< 0.01, ***P<0.001, **** P<0.0001). Number of patients with gene expression data for each immune subtype is detailed below the bar plots.

4.7 Differential gene expression signatures between Wound Healing and IFN-γ Dominant subtypes identify key genes and pathways of potential diagnostic or therapeutic relevance in CRC patients

We analysed gene expression profiles of the 2 major ISs in CRC (Wound Healing and IFN-γ Dominant subtypes), both in the global cohort of CRC patients and by CMS to address the interplay of both classifications. 3150 and 1362 differentially expressed genes achieved statistical significance (*FDR< 0.05*) when we analysed all and CMS1 CRC patients, respectively (Figure 9A and Supplementary results 1-2). 239, 108 and 327 genes showed significant differential expression when these subtypes (Wound Healing and IFN-γ Dominant subtypes) were compared within the CMS2, CMS3 and CMS4 groups, respectively (Figure 9A and Supplementary results 3-5). As shown in Figure 9, the expression profile of an important number of genes was specific to each CMS, but 21 genes were consistently found to be differentially expressed in all analysis performed (Figure 9A and Supplementary Table 13).

Then, we analysed enrichment score gensets to determine the upregulation of relevant pathways within immune subtypes. Analysis of the whole cohort showed that the Wound Healing subtype was enriched in metabolic pathways and had greater activation of Wnt and hedgehog signalling (Figure 9B). In contrast, the IFN-γ Dominant subtype presented greater activation of pathways related to the immune system, apoptosis and

DNA repair, as well as mTOR signalling and oxidative phosphorylation (Figure 9B and Supplementary Table 14). Similar results were found when CMS1 group was analysed, being metabolism related pathways such as glycolysis and pyruvate metabolism overrepresented in the Wound Healing subtype, and apoptosis and reactive oxygen species pathways characteristic of the IFN-γ Dominant group (Supplementary Figure 12 and Supplementary Tables 15-18).

This result suggests that clear biological differences exist between IFN-γ Dominant and Wound Healing samples apart from immune system activation. Some of these differences such as metabolism, matrix remodeling and other pathways in cancer could explain the distinct immune phenotype.

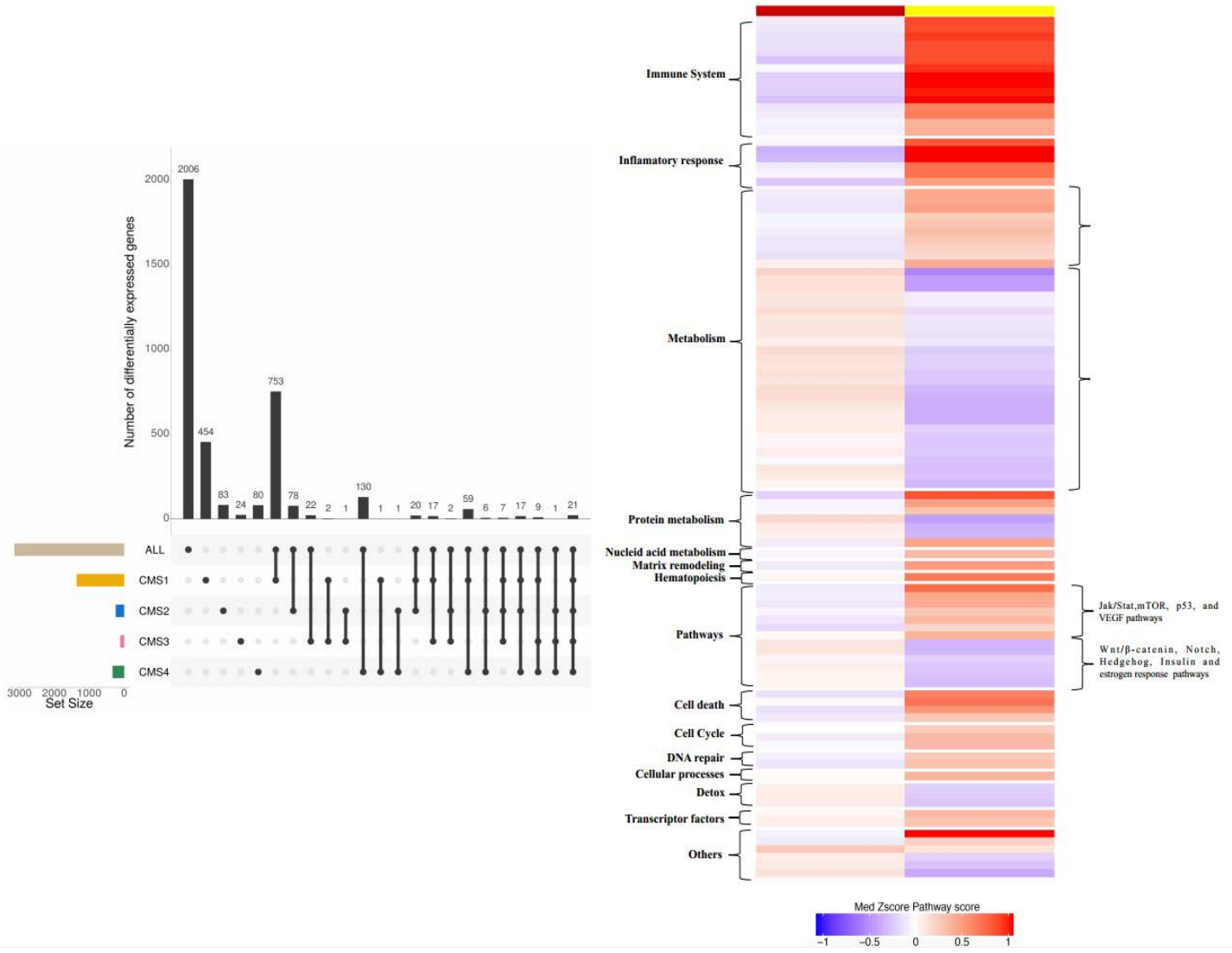


Figure 9: Genes and pathways differentially expressed between wound healing and IFN-γ dominant immune subtypes. A) Upset plot showing differentially expressed genes between wound healing and IFN-γ dominant immune subtypes in all cohort and in each CMS subtype. B) Heatmap showing the Gene set mRNA enrichment analysis (ssGSEA) of signatures (median z - scores) of special interest in CRC among the 2 predominant immune subtypes, wound healing and IFN-γ dominant.

5. Discussion

An increasing body of evidence supports the major role that tumor microenvironment and, in particular, the immune system plays in cancer fate[52]. An improved understanding of the immune landscape of tumors is, therefore, critical to refine immunotherapeutic strategies that have actually revolutionized cancer care but are still ineffective in a great proportion of CRC patients. Besides, nowadays only few molecular biomarkers are considered and in many cases are used separately. The advances in the sequencing technology will provide more accurate expression data, that have the potential to describe the microenvironment status using many information together. In this way, Thorsson *et al.* have described six immune subtypes that encompass nearly all human malignancies, with distinct immunogenomic features and a significant impact in clinical outcome[16]. Nevertheless, the distribution of these immune subtypes substantially differed across the 33 solid tumors included, and it seems that their association with patients' prognosis was also dependent on tumor type.

In the present study, we have specifically characterized the immune subtypes in the TCGA cohort of CRC patients, including clinical and pathological features, genomic and transcriptomic profiles, and composition and functional orientation of the immune and stromal cell components of tumor microenvironment. In our study we show that among CRC patients only 5 of the 6 immune subtypes are present, with 2 predominant ISs: the C1 Wound Healing subtype, identified in 459 samples (77%), and the C2 IFN-γ Dominant subtype, present in 103 samples (17%). The other 3 ISs were far less commonly encountered in CRC (C3 Inflammatory: 18 samples, 3%; C4 Lymphocyte Depleted: 13 samples, 2%; and C6 TGF-β Dominant: 4 samples, 0.7%). This result suggests that there is not a lot heterogeneity in an immune point of view in colorectal cancer and only a 23%

of patients belong to another group that is not the principal IS. Nevertheless, the fact that this classification is made in several solid tumors could dilute existing differences in this specific tissue but at the same time it makes the groups more robust and, in theory, with more remarkable differences among them.

We have also studied the interplay between IS classification and CMS classification[15] . We could see that these both classification systems had related each to other with clearly enrichments, but do not overlap completely. Indeed, the C1 Wound Healing IS was particularly dominant in CMS2 (91%) but far less common in CMS1 (46%) tumors, whereas the proportion of C1 Wound Healing in CMS3 and CMS4 tumors was similar to the average observed for the overall CRC population (77-78%). On the contrary, the C2 IFN-γ Dominant was the most common IS observed in CMS1 tumors (53%) and was under-represented in CMS2 tumors (8%), with a slightly higher representation in CMS3 and CMS4 tumors (11-13%). Other ISs were barely present in CMS1 and CMS2 tumors, whereas the immunological landscape of CMS3 and CMS4 was more diverse: they had some representation of the C3 Inflammatory (7% and 6%, respectively) and C4 Lymphocyte Depleted (4.3% and 1.5%, respectively) subtypes, and all 3 cases with the TGF-β Dominant phenotype belonged exclusively to the CMS4 subgroup (2.3%) . This result could contrast with the idea that CMSs are homogenous groups in an immune point of view with a potential predictor power to immunotherapy response[2,7,33,53–55]. Thus, it leads us to wonder what information could provide IS classification to the well-known CMS classification.

First, we looked for enrichments in several molecular and clinical features in the different IS, founding some of them with significance. However, this significance disappeared in the analysis made in each subtype. This suggest that differences in the

whole cohort among IS are provided by the enrichment in the distinct CMS but not by itself. Nevertheless, it seems to have importance in the prognosis when Cox multivariate regression showed that the IS classification was significantly associated with survival in CRC patients, independent of other well-established prognostic factors including age, stage at diagnosis and primary tumor site, whereas the CMS classification lost statistical significance in the multivariate analysis. These observations could be consistent with the fact that biological differences were found among immune subtypes, even inside each CMS subtype. For instance, CMS1(MSI-like Immune) showed clear dissimilarities between C1 wound healing subtype and C2 IFN-γ dominant subtype. To date, CMS1 patients had been postulated to be the most likely to benefit from immune-checkpoint inhibitors, as this subtype includes most MSI tumors, which are highly mutated tumors due to deficient DNA mismatch repair mechanisms[56,57]. Nevertheless, not all CMS1 are MSI tumors and other CMS subgroups (i.e, CMS3) also include a significant proportion of MSI patients. This is relevant as MSI is the biomarker used in pivotal studies that demonstrated the clinical efficacy of checkpoint inhibitors in CRC, and the only available biomarker in standard clinical practice to identify immunogenic CRC[17–20,23]. Indeed, MSI has been the first biomarker to define a Food and Drug Administration (FDA) approval of an anticancer drug (pembrolizumab) for a tumor/site-agnostic indication[58]. Moreover, our results reported that the immunotherapy landscape of CMS1 is diverse and this may potentially impact response to therapy. Given the fact that IFN-γ Dominant tumors presented a high expression of CD8+, follicular helper T cells and M1 macrophages, as well as high levels of regulatory T cells, M2 macrophages, dendritic cells and neutrophils, and this IS was also characterized by the upregulation of several immune regulatory genes (PD-1/PD-L1, CTLA4, IDO1 and LAG3) likely responsible for its immune surveillance escape mechanism, we could hypothesize that the IFN-γ Dominant IS may be responsible

in part of the heavy immune activation and susceptibility of CMS1 tumors to immune-checkpoint inhibition, whereas CMS1 tumors of the Wound Healing subtype may likely be less immunogenic and more resistant to these agents and require other therapeutic approaches. In addition, IFN-γ Dominant tumors belonging to CMS2-4 subgroups could also potentially benefit of immune-checkpoint targeted therapy. This is particularly relevant as MSI tumors only represent a small proportion of patients with advanced CRC (~5%), and despite the unquestionable success of immunotherapy in this subgroup of patients, still a significant proportion of MSI tumors (40-60%) do not respond to checkpoint inhibitors for yet unexplored biological grounds. On the contrary, a subset of MSS tumors also show increased expression of immune genes, indicating that other factors also determine immune infiltration and clinical outcome. In fact, some immune profiles such as the CRC immunescore[59], based on the histological quantitation and localization of cytotoxic and memory T cells in the tumor center and invasive margin, proved to be a stronger predictor of patient survival than MSI[60]. The new IS classification could be therefore a new valuable tool to aid in the selection of patients for immune therapy.

Another relevant finding of our study was that the prognostic impact of immune subtypes in CRC patients was different from that reported for the global TCGA dataset[16]. Considering the two most prevalent IS in CRC, it is to be noted that differences in survival between the Wound Healing and IFN-γ Dominant subtypes were more pronounced as compared to the overall cancer population including all solid tumors. IFN-γ is a cytokine mainly produced by activated T lymphocytes and by natural killer cells in response to a variety of immune stimuli. The activation of IFN-γ pathway directly results in the inhibition of tumor cell growth and the improvement of immune response, which recognizes and eliminates tumor cells. Specifically in CRC, IFN-γ induces tumor cell

cycle arrest by enhancing expression of p27, p16 or p21[61] and autophagy-associated apoptosis[62]. However, the pro-tumoral functions of IFN-γ involve proliferative and anti-apoptotic signals, as well as cancer evasion by promoting angiogenesis or the expression of immune-checkpoints[63,64]. Particularly, several reports have demonstrated that IFN-γ induce the expression of PDL-1, IDO and CTL-4 in tumor cells and were associated with poor prognosis in several cancer types[63].

Some of these observations may be partially conditioned by some relevant caveats inherent to the use of TCGA data. First, the fact that for most tumor types a tumor cell component > 50% was required for study entry; this introduces a significant bias as the epithelial cell component is likely over-represented and the most immune-infiltrated tumors are excluded from the analysis. In addition, survival rates and patient follow-up substantially differed across tumor types, antitumor therapy is presumably very heterogeneous in the study population, and its impact on clinical outcome is not considered.

Finally, it should be pointed out that the IS classification was developed in a large but very heterogeneous cancer population, which may dilute distinct transcriptomics profiles within each tumor type with potentially relevant clinical implications. This is also reflected in the lack of high variance in immune subtypes between CMS4 and CMS3 when immunogenic differences were found in previous publications[33]. Thus, further refinement of this classification specifically adjusted for CRC is certainly warranted. Moreover, we did not find the same biologic differences between Wound healing and IFN-γ dominant in all CMS. Starting by immune population, we could see that IFN-γ dominant in CMS1 is enriched in several immune populations while in other subtypes it is not happen, despite of in all CMS are differences in the expression of interferon gamma

and alpha between the principal ISs. Furthermore, not all genes are differentially expressed between Wound Healing and IFN-γ dominant in all CMS, existing genes only found in one of them. We are probably facing a different molecular scenario where IFN-γ is differentially activated in all of them, but with distinct biological meanings. Besides, is interesting to study what could be underlying these differences in the expression of IFN-γ. The presence of T cells and Natural killer cells populations infiltrated in the tumor has been recently associated with the TMB, which increase the probability to a neoantigen would be recognized promoting the expression of cytokines[64]. However, we did not see such variations inside each CMS cohort, what means that other processes are involved. For instance, some of the genes differentially expressed between Wound Healing and IFN-γ Dominant in CMS1 are related to proteasome and immunoproteasome (PSME2 FDR = 0.01 E-4; PSME1 FDR= 0.005; PSMB8; FDR = 0.01; PSMB10 FDR= 0.01; PSMA4 FDR = 0.01; PSMB9 FDR = 0.02; PSMA6; FDR= 0.02; PSMA1 FDR = 0.043) (Supplementary results 2). Immunoproteasome is necessary to process mutated proteins in peptides to be presented so the incorrect function of this complex do not allow that peptides could be presented by molecular histocompatibility complex (MHC). The lack of IFN-γ in Wound Healing tumors could be the cause of this down expression, since it is well known that this is the main cytokine which promote proteasome related genes. Furthermore, we could find overexpressed genes in Wound Healing samples that inhibit different aspects of immune response (GP2 FDR = 6E-6; FGL1 FDR = 5E-4; SPINK4 FDR = 0.04) (Supplementary results 2). GP2 reduced innate and adaptive immune responses at several levels, FGL1 is a known ligand of LAG3 promoting the inhibition of T cells and SPINK4 that encodes for a serine peptidase inhibitor that has been associated to inhibit granzymes cytotoxic effect used by CD8+ T and NK cells[65–67]. On the other hand, CMS2 Wound healing samples have an overexpression of LNX1 (FDR = 8.26E-5)

which encodes for membrane protein that binds CD8+ T cell protein provoking it is ubiquitination and degradation what result in the incorrect function of the cell[68](Supplementary results 3).

In summary, in line with the changing treatment paradigm, that is shifting from the traditional one predominantly focused on targeting the epithelial compartment, to the development of more integrated approaches targeting tumor microenvironment, in-depth study of the immune landscape of tumors provides very valuable information for cancer management.

6. Conclusions

In the present study we have characterized the recently described pan-cancer IS classification in a large cohort of CRC patients, demonstrating distinct clinical and biological implications of ISs in this cancer type. We have also identified substantial heterogeneity in the distribution of ISs by CMS subgroups. Moreover, we show that IS classification identify tumors which have different microenvironments, although belong to the same CMS, being interesting to highlight that we can separate the immune molecular subtype CMS1 in two different IS with clear immune differences. It is expected that profound biological differences observed among ISs and CMS translate into heterogeneous drug responses, both to conventional cytotoxic drugs and alternative treatment strategies targeting the tumor ecosystem, including immunotherapy. However, this result should be validated in an independent cohort of patients and the clinical implications must be tested in clinical trials.

7. References

1. Basile, D. *et al.* Immunotherapy for colorectal cancer: where are we heading? *Expert Opinion on Biological Therapy* **17**, 709–721 (2017).

2. Ciardiello, D. *et al.* Immunotherapy of colorectal cancer: Challenges for therapeutic efficacy. *Cancer Treatment Reviews* **76**, 22–32 (2019).

3. Lièvre, A. *et al.* Protein biomarkers predictive for response to anti-EGFR treatment in RAS wild-type metastatic colorectal carcinoma. *British Journal of Cancer* **117**, 1819–1827 (2017).

4. Ganesh, K. *et al.* Immunotherapy in colorectal cancer: rationale, challenges and potential. *Nature Reviews Gastroenterology & Hepatology* **16**, 361–375 (2019).

5. Kalyan, A., Kircher, S., Shah, H., Mulcahy, M. & Benson, A. Updates on immunotherapy for colorectal cancer. *Journal of Gastrointestinal Oncology* **9**, 160–169 (2018).

6. Garrido-Castro, A. C., Lin, N. U. & Polyak, K. Insights into Molecular Classifications of Triple-Negative Breast Cancer: Improving Patient Selection for Treatment. *Cancer Discovery* **9**, 176–198 (2019).

7. Dienstmann, R. *et al.* Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature Reviews Cancer* **17**, 79–92 (2017).

8. Collisson, E. A., Bailey, P., Chang, D. K. & Biankin, A. V. Molecular subtypes of pancreatic cancer. *Nature Reviews Gastroenterology & Hepatology* **16**, 207–220 (2019).

9. Budinska, E. *et al.* Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer: Gene expression heterogeneity in colorectal cancer. *The Journal of Pathology* **231**, 63–76 (2013).

10. De Sousa E Melo, F. *et al.* Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Medicine* **19**, 614–618 (2013).

11. Marisa, L. *et al.* Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *PLoS Medicine* **10**, e1001453 (2013).

12. Roepman, P. *et al.* Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition: Molecular subtypes in colorectal cancer. *International Journal of Cancer* **134**, 552–562 (2014).

13. Sadanandam, A. *et al.* A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine* **19**, 619–625 (2013).

14. Schlicker, A. *et al.* Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Medical Genomics* **5**, (2012).

15. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**, 1350–1356 (2015).

16. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830.e14 (2018).

17. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine* **372**, 2509–2520 (2015).

18. Imran, A. *et al.* Role of Molecular Biology in Cancer Treatment: A Review Article. *Iran. J. Public Health* **46**, 1475–1485 (2017).

19. Jasperson, K. W., Tuohy, T. M., Neklason, D. W. & Burt, R. W. Hereditary and Familial Colon Cancer. *Gastroenterology* **138**, 2044–2058 (2010).

20. *Cancer Control Opportunities in Low- and Middle-Income Countries*. (National Academies Press, 2007). doi:10.17226/11797

21. The Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **45**, 1113–1120 (2013).

22. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific Reports* **3**, (2013).

23. Zhu, K. *et al.* Oncogenes and tumor suppressor genes: comparative genomics and network perspectives. *BMC Genomics* **16 Suppl 7**, S8 (2015).

24. Siegel, R. L. *et al.* Colorectal cancer statistics, 2017: Colorectal Cancer Statistics, 2017. *CA: A Cancer Journal for Clinicians* **67**, 177–193 (2017).

25. Ren, Z. & Tao, Z. Molecular Basis of Colorectal Cancer: Tumor Biology. in *Surgical Treatment of Colorectal Cancer* (eds. Kim, N. K., Sugihara, K. & Liang, J.-T.) 23–34 (Springer Singapore, 2018). doi:10.1007/978-981-10-5143-2_3

26. Terry, P. *et al.* Fruit, Vegetables, Dietary Fiber, and Risk of Colorectal Cancer. *JNCI Journal of the National Cancer Institute* **93**, 525–533 (2001).

27. Han, S. *et al.* Role of intestinal flora in colorectal cancer from the metabolite perspective: a systematic review. *Cancer Management and Research* **Volume 10**, 199–206 (2018).

28. Morán, A. Differential colorectal carcinogenesis: Molecular basis and clinical relevance. *World Journal of Gastrointestinal Oncology* **2**, 151 (2010).

29. Oiseth, S. J. & Aziz, M. S. Cancer immunotherapy: a brief review of the history, possibilities, and challenges ahead. *Journal of Cancer Metastasis and Treatment* **3**, 250 (2017).

30. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLOS Computational Biology* **13**, e1005457 (2017).

31. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE* **9**, e78644 (2014).

32. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

33. Becht, E. *et al.* Immune and Stromal Classification of Colorectal Cancer Is Associated with Molecular Subtypes and Relevant for Precision Immunotherapy. *Clinical Cancer Research* **22**, 4057–4066 (2016).

34. Roma-Rodrigues, C., Mendes, R., Baptista, P. & Fernandes, A. Targeting Tumor Microenvironment for Cancer Therapy. *International Journal of Molecular Sciences* **20**, 840 (2019).

35. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

36. Goldman, M. *et al.* The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. *bioRxiv* (2019). doi:10.1101/326470

37. Gao, J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science Signaling* **6**, pl1–pl1 (2013).

38. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

39. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology* **12**, R41 (2011).

40. R Development Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2008).

41. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53**, 457 (1958).

42. Peto, R. *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *British Journal of Cancer* **35**, 1–39 (1977).

43. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods* **12**, 453–457 (2015).

44. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

45. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, (2014).

46. Yoav Benjamini, Y. H. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Royal Stat. Soc.* **57**, 289–300 (1995).

47. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).

48. Carlson, M. *KEGG.db: A set of annotation maps for KEGG*. (2016).

49. Gehlenborg, N. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. (2019).

50. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, 7 (2013).

51. Hervé, M. *RVAideMemoire: Testing and Plotting Procedures for Biostatistics*. (2019).

52. Wang, M. *et al.* Role of tumor microenvironment in tumorigenesis. *Journal of Cancer* **8**, 761–773 (2017).

53. Sveen, A. *et al.* Colorectal Cancer Consensus Molecular Subtypes Translated to Preclinical Models Uncover Potentially Targetable Cancer Cell Dependencies. *Clinical Cancer Research* **24**, 794–806 (2018).

54. Spallanzani, A. *et al.* Immunotherapy in the treatment of colorectal cancer: a new kid on the block. *Journal of Cancer Metastasis and Treatment* **4**, 28 (2018).

55. Roelands, J. *et al.* Immunogenomic Classification of Colorectal Cancer and Therapeutic Implications. *International Journal of Molecular Sciences* **18**, 2229 (2017).

56. Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).

57. Gelsomino, F., Barbolini, M., Spallanzani, A., Pugliese, G. & Cascinu, S. The evolving role of microsatellite instability in colorectal cancer: A review. *Cancer Treatment Reviews* **51**, 19–26 (2016).

58. Lemery, S., Keegan, P. & Pazdur, R. First FDA Approval Agnostic of Cancer Site — When a Biomarker Defines the Indication. *New England Journal of Medicine* **377**, 1409–1412 (2017).

59. Galon, J. *et al.* Cancer classification using the Immunoscore: a worldwide task force. *Journal of Translational Medicine* **10**, (2012).

60. Mlecnik, B. *et al.* Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability. *Immunity* **44**, 698–711 (2016).

61. Wang, L., Wang, Y., Song, Z., Chu, J. & Qu, X. Deficiency of Interferon-Gamma or Its Receptor Promotes Colorectal Cancer Development. *Journal of Interferon & Cytokine Research* **35**, 273–280 (2015).

62. Wang, Q.-S., Shen, S.-Q., Sun, H.-W., Xing, Z.-X. & Yang, H.-L. Interferon-gamma induces autophagy-associated apoptosis through induction of cPLA2-dependent mitochondrial ROS generation in colorectal cancer cells. *Biochemical and Biophysical Research Communications* **498**, 1058–1065 (2018).

63. Seidel, J. A., Otsuka, A. & Kabashima, K. Anti-PD-1 and Anti-CTLA-4 Therapies in Cancer: Mechanisms of Action, Efficacy, and Limitations. *Frontiers in Oncology* **8**, (2018).

64. McGrail, D. J. *et al.* Multi-omics analysis reveals neoantigen-independent immune cell infiltration in copy-number driven cancers. *Nature Communications* **9**, (2018).

65. Werner, L. *et al.* Identification of Pancreatic Glycoprotein 2 as an Endogenous Immunomodulator of Innate and Adaptive Immune Responses. *The Journal of Immunology* **189**, 2774–2783 (2012).

66. Xie, M., Li, K., Li, J., Lu, D. & Hu, B. Association and diagnostic value of serum SPINK4 in colorectal cancer. *PeerJ* **7**, e6679 (2019).

67. Wang, J. *et al.* Fibrinogen-like Protein 1 Is a Major Immune Inhibitory Ligand of LAG-3. *Cell* **176**, 334-347.e12 (2019).

68. D'Agostino, M. *et al.* Ligand of Numb proteins LNX1p80 and LNX2 interact with the human glycoprotein CD8 and promote its ubiquitylation and endocytosis. *Journal of Cell Science* **124**, 3545–3556 (2011).